Paper Title: Knowledge Transfer via Multiple Model Local Structure Mapping

End Term Review
“This work is done as part of IE 506 Course Project”

Team Name: *Mayur_B*

Presented by:
Name: *Mayur Ramesh Bhurle*
Roll Number: *23M1027*

Guided By: *Prof. Balamurugan Palaniappan*

Outline

1. Problem Addressed
2. Summary of work done before Mid-term project review
3. Major comments given during the mid-term project review
4. Data Description
5. Confidence Based Ensemble
6. Results Obtained
7. Variation in accuracy with Hyperparameter (p)
8. Reasons for Variations in Results
9. References

1. Problem Addressed

The paper addresses the problem of transfer learning, where the goal is to learn from one or several training domains and make predictions in a different but related test domain.

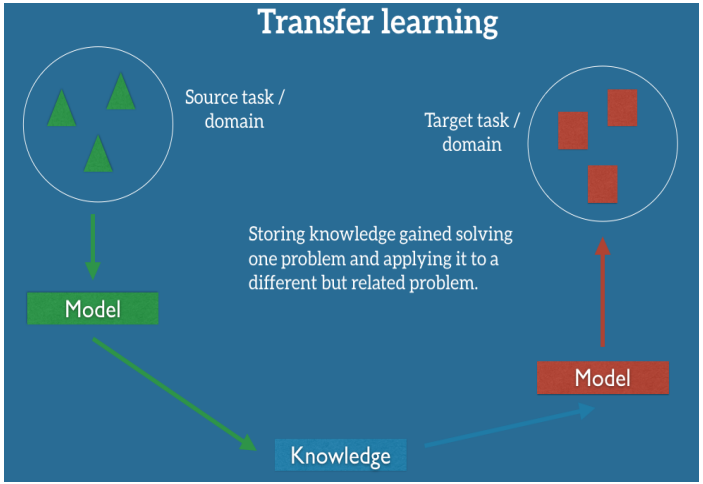


Fig 1: Transfer Learning Concept

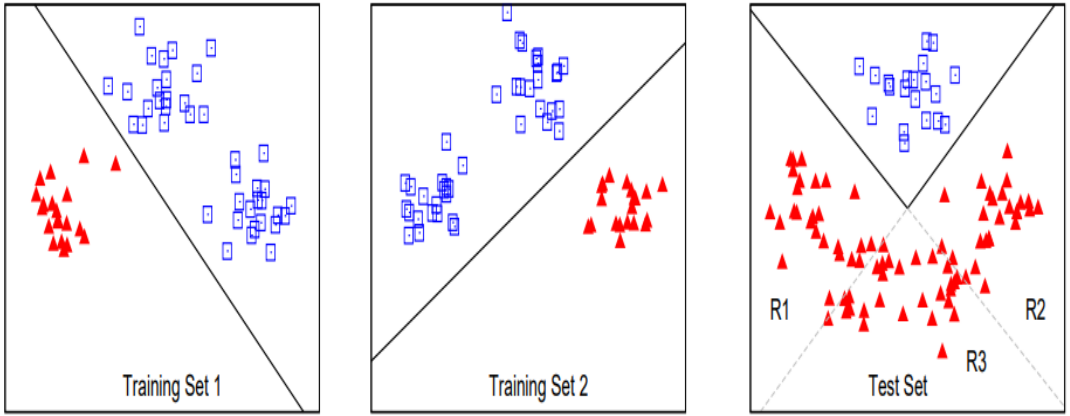


Fig 2: Transfer Learning example

(Source: Transfer Learning - Machine Learning's Next Frontier, Sebastian Ruder, accessed on 04.05.2024 (https://www.ruder.io/transfer-learning/)

2. Summary of Work Done Before Mid-Term Project Review

Step 1: Graph Based Weight Estimation:

$$w_{M,\mathbf{x}} \propto s(G_M, G_T; \mathbf{x}) = \frac{\sum_{v_1 \in V_M} \sum_{v_2 \in V_T} \mathbf{1}\{v_1 = v_2\}}{|V_M| + |V_T|} \quad \text{..... (1)}$$

$$w_{M_i,\mathbf{x}} = \frac{s(G_{M_i}, G_T; \mathbf{x})}{\sum_{i=1}^{k} s(G_{M_i}, G_T; \mathbf{x})}, \quad \text{..... (2)}$$

$$P(y|E, \mathbf{x}) = \sum_{i=1}^{k} w_{M_i,\mathbf{x}} P(y|M_i, \mathbf{x}), \quad \text{..... (3)}$$

$$y^* = \arg\max_y P(y|E, \mathbf{x}). \quad \text{..... (4)}$$

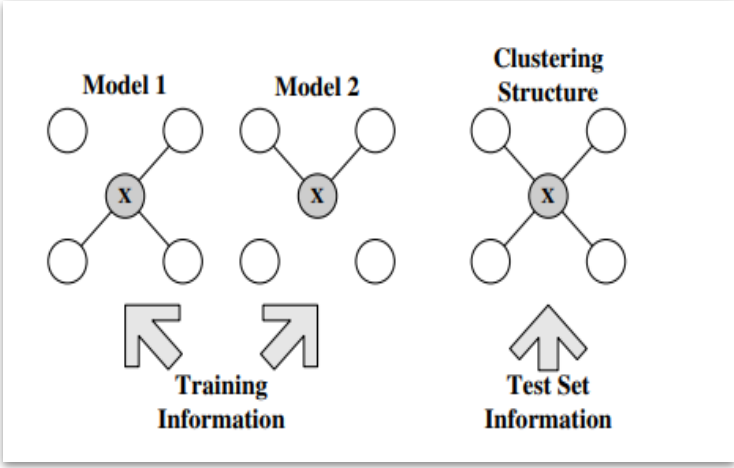


Figure 2: Local Neighborhood Graphs around x

2. Summary of Work Done Before Mid-Term Project Review

Step 2: Local Structure Based Adjustment:

$$s_{avg}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} s(G_{M_i}, G_T; v) \quad \text{..... (5)}$$

- *Savg(x)* ≥ δ, where δ is the threshold, we consider prediction obtained from Eq. (3), that is
- The “unsupervised” classifier U is not trained on any labeled training set. Its prediction on x is mainly determined by the neighbors of x with labels predicted by the combined classifier.

$$P(y|U, \mathbf{x} \in C) \approx \frac{P(y, \mathbf{x} \in C'|E)}{P(\mathbf{x} \in C')} \approx \frac{c(y, C'|E)}{|C'|} \quad \text{..... (6)}$$

- where c(y, C’ | E) is the number of examples with label y predicted by ensemble E in C’.

2. Summary of Work Done Before Mid-Term Project Review

Comparing Accuracy & MSE with and without PCA (for “R vs T” case, 20 newsgroup dataset)

Algorithm	Accuracy (with PCA)	MSE (with PCA)
Decision Tree	0.640	0.360
Logistic Reg.	0.623	0.377
SVM	0.659	0.341
SMA	0.6338	0.3661
LS-SVM	0.2895	0.7104
p-LWE	0.6296	0.3703
LWE	0.7563	0.2536

Algorithm	Accuracy (without PCA)	MSE (without PCA)
Decision Tree	.631	0.369
Logistic Reg.	0.619	.0318
SVM	0.659	0.341
SMA	0.608	0.391
LS-SVM	0.276	0.723
p-LWE	0.624	0.375
LWE	0.759	0.240

3. Major comments given during the mid-term project review:

For the final presentation the team can attempt:

1. The team can try without PCA and check the results and get back.
2. Time series/financial data (Should be tried by using appropriate methods for classification and clustering)
3. Can try weighing scheme in the algorithm.
4. Intrusion detection dataset can be tried.
5. Presentation should be made more enthusiastic

4. Data Description

Features considered:

1. **VWAP** - Volume-Weighted Average Price: VWAP is a ratio of the cumulative share price to the cumulative volume traded over a given time period. It provides insight into the average price at which a stock is traded, weighted by the volume of trades.
2. **Volume** - Volume Traded for the Day: Represents the total number of shares or contracts traded during a specific time period, typically a trading day.
3. **Turnover** - Turnover Ratio: The turnover ratio is the ratio of sellers to buyers of a stock. It helps in understanding the market activity and liquidity.

1. **Deliverable Volume** - Amount of Deliverable Volume: Represents the volume of shares that were actually delivered (transferred) as opposed to being traded intraday.

4. Data Description

Features capturing Temporal dependencies:

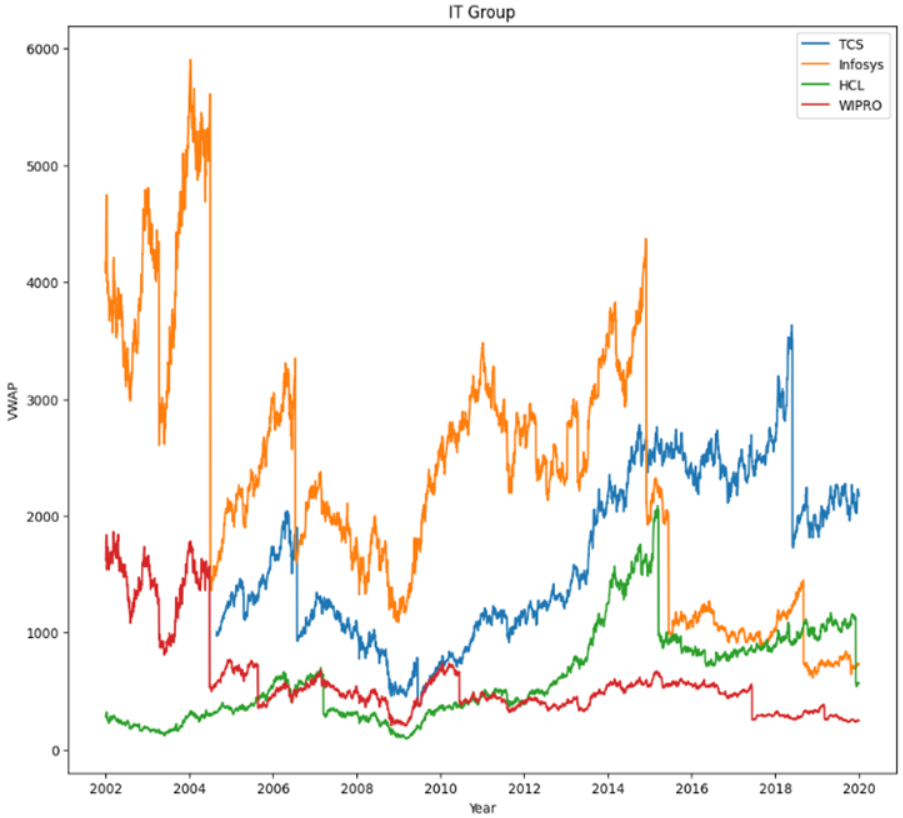
1. Month/Day/Date
2. Lag (no. of lags = 3)
3. Rolling Mean (window size =5)
4. Rolling Min
5. Rolling Max

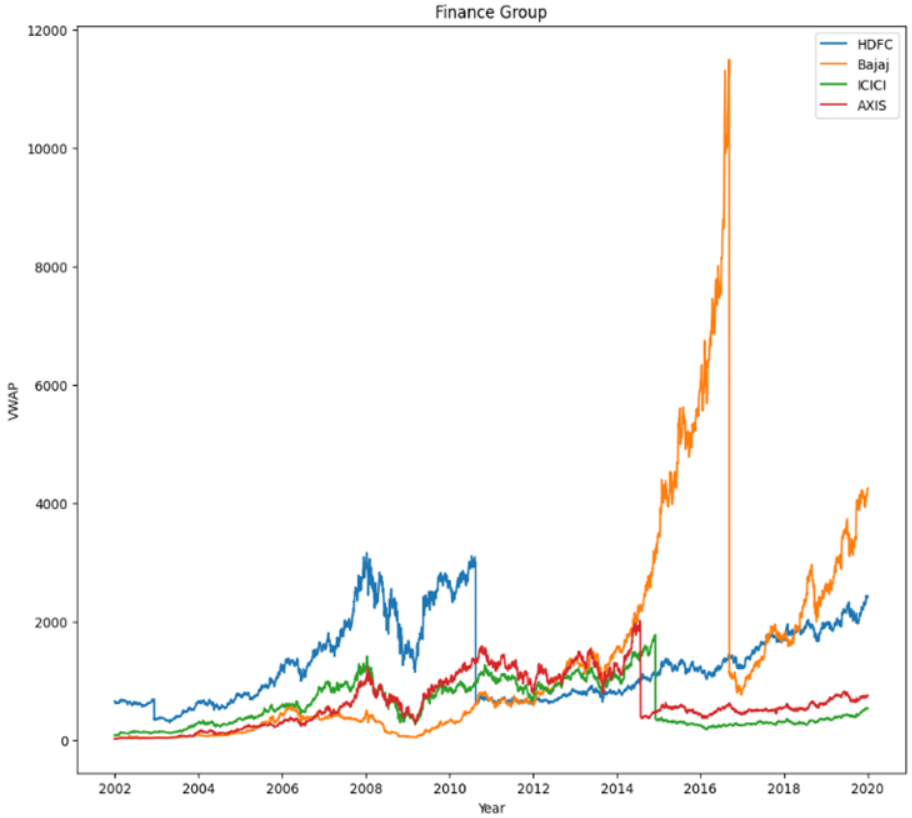
Stock features

1. VWAP(Volume-Weighted Average Price)
2. Volume
3. Turnover
4. Deliverable Volume

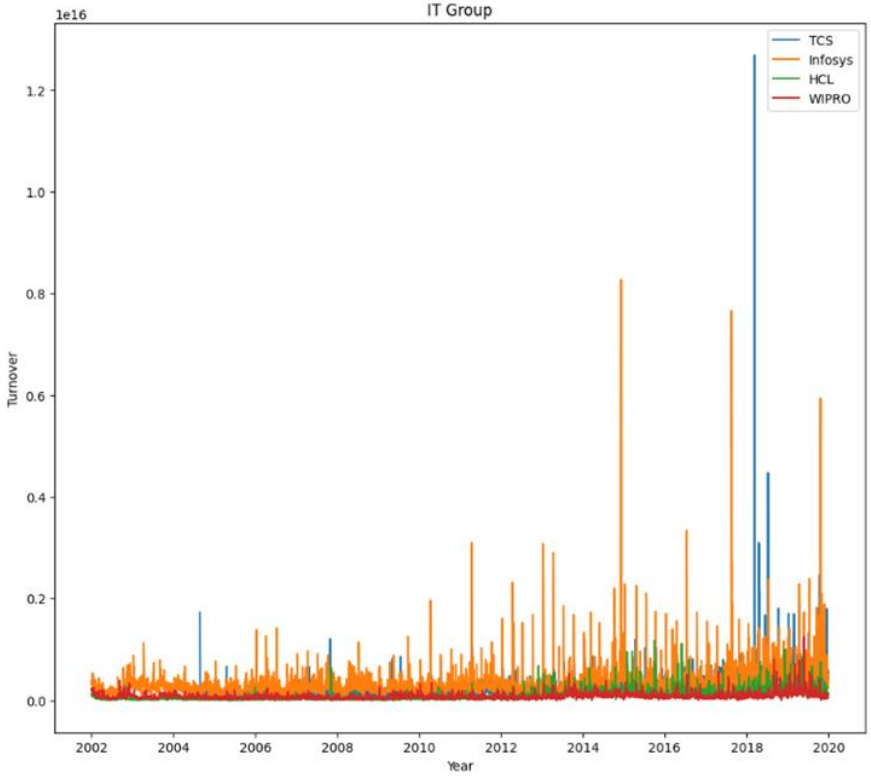
	Date	Prev Close	VWAP	Volume	Turnover	Deliverable Volume	Year	Label	Month	Day	...	lag_2	lag_3	rolling_mean	rolling_min	rolling_max	rolling_std	rolling_mean1	rolling_min1	rolling_max1	rolling_std1
502	2002-01-07	643.00	624.93	96917	6.060000e+12	81153.0	2002	0	1	7	...	649.79	653.85	646.908	624.93	661.89	13.892333	25328.6	5604.0	81153.0	31408.369972
503	2002-01-08	624.95	623.10	76678	4.780000e+12	65568.0	2002	0	1	8	...	644.08	649.79	639.150	623.10	653.85	14.260219	37321.4	11897.0	81153.0	33380.218787
504	2002-01-09	620.55	640.78	107845	6.910000e+12	84360.0	2002	0	1	9	...	624.93	644.08	636.536	623.10	649.79	11.893457	51647.0	11897.0	84360.0	35492.369765
505	2002-01-10	644.55	646.96	21653	1.400000e+12	16928.0	2002	0	1	10	...	623.10	624.93	635.970	623.10	646.96	11.149067	51981.2	11897.0	84360.0	35069.401402
506	2002-01-11	647.05	647.66	98620	6.390000e+12	79018.0	2002	0	1	11	...	640.78	623.10	636.686	623.10	647.66	11.890411	65405.4	16928.0	84360.0	28029.678000

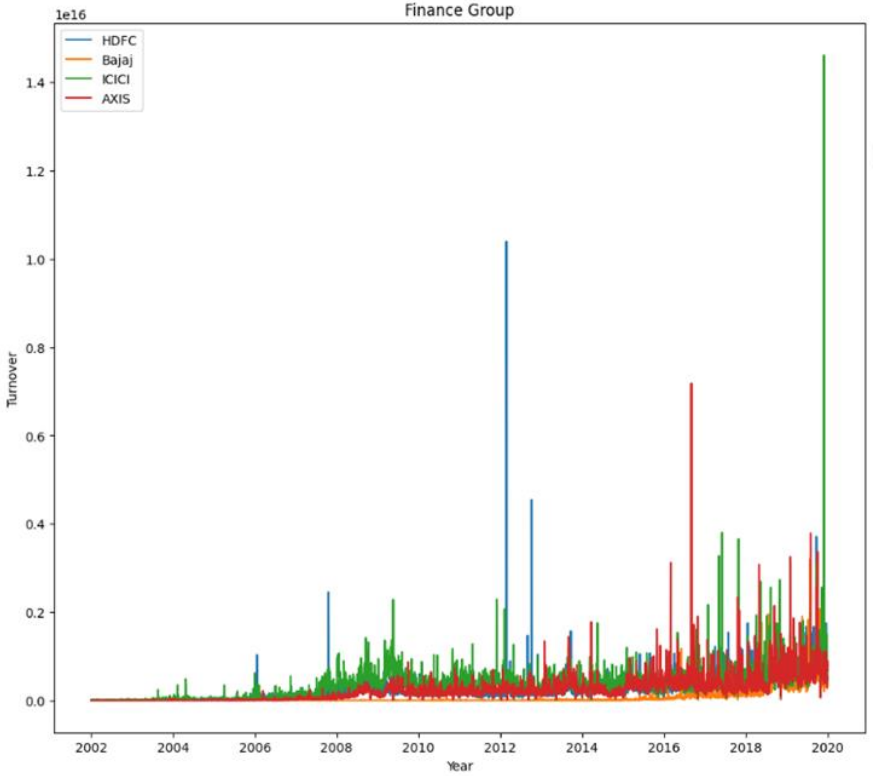
Stock Price Data (VWAP : Volume-Weighted Average Price)



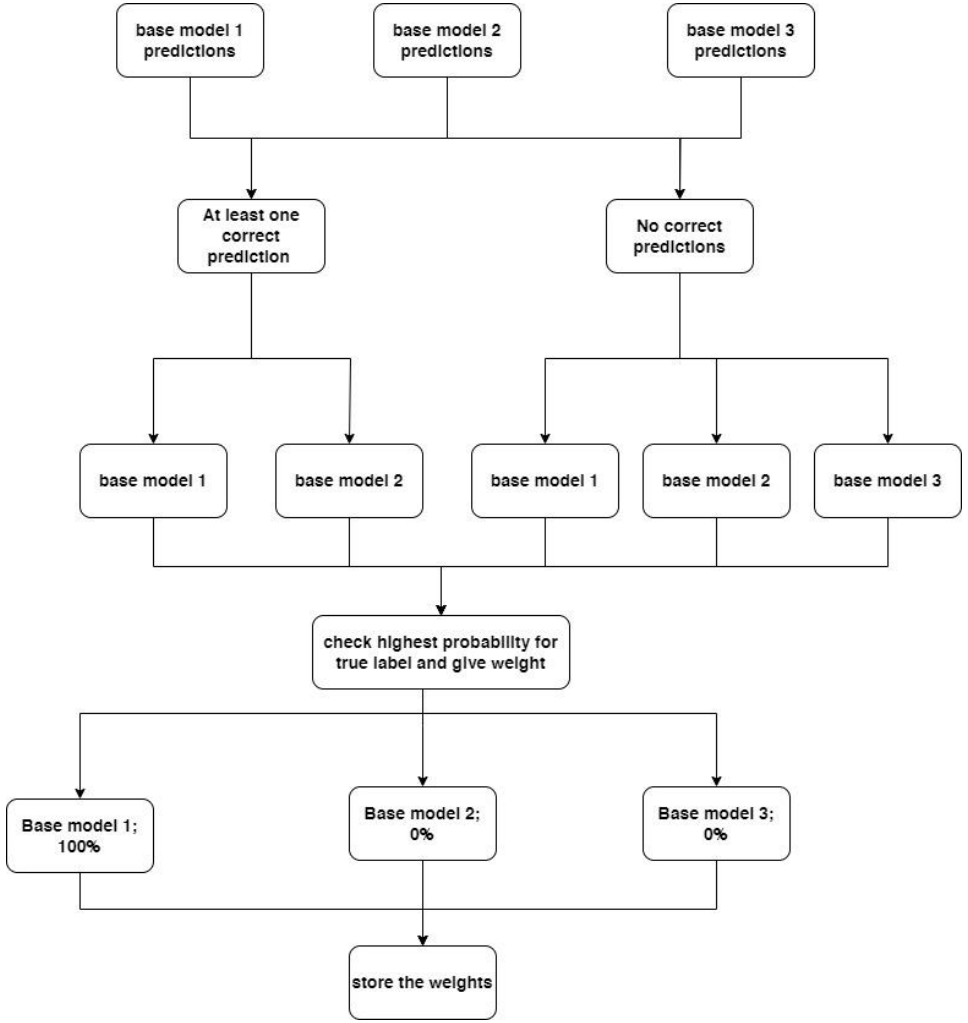


Stock Price Data (Turnover ratio)





5. Confidence based Ensemble



6. Results Obtained

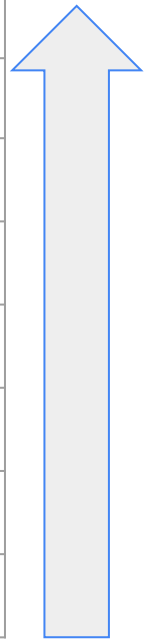
Sr. No.	*Algorithm*	*Accuracy*		
		IT vs Finance	**IT vs Energy**	**Finance vs Energy**
1	**LSTM**	0.6154	.3290	0.4453
2	**SVM**	0.3088	0.1491	0.2918
3	**LR**	0.4761	0.1834	0.2763
4	**CNN**	0.3731	0.2625	0.4605
6	**pLWE**	0.4748	0.1327	0.5
7	**LWE**	0.4590	0.3331	0.5
8	**Confidence Based Ensemble**	0.7698	0.4052	0.6102

6. Results Obtained

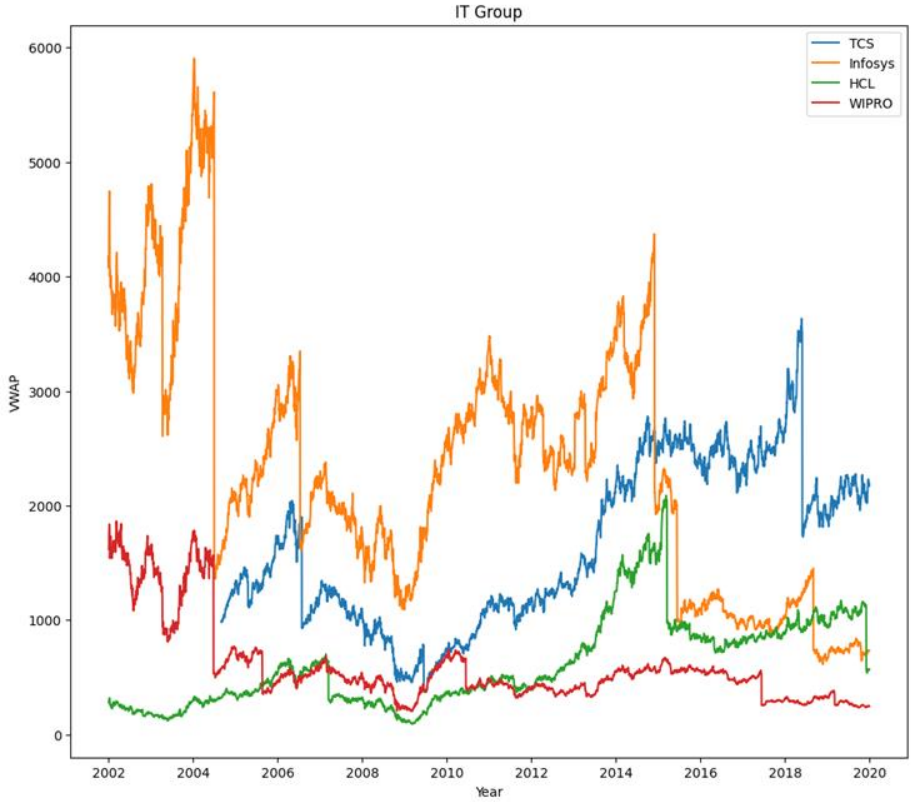
Sr. No.	***Algorithm***	***MSE***		
		IT vs Finance (Bajaj & TCS)	**IT vs Energy** (BPCL & TCS)	**Finance vs Energy** (Bajaj and BPCL)
1	**LSTM**	0.3845	0.6709	0.5546
2	**SVM**	0.6911	0.8508	0.7081
3	**LR**	0.5238	0.8165	0.7836
4	**CNN**	0.6268	0.7374	0.5394
6	**pLWE**	0.2993	0.4752	0.5
7	**LWE**	0.5409	06668	0.5
8	**Confidence Based Ensemble**	0.2301	0.5947	0.3897

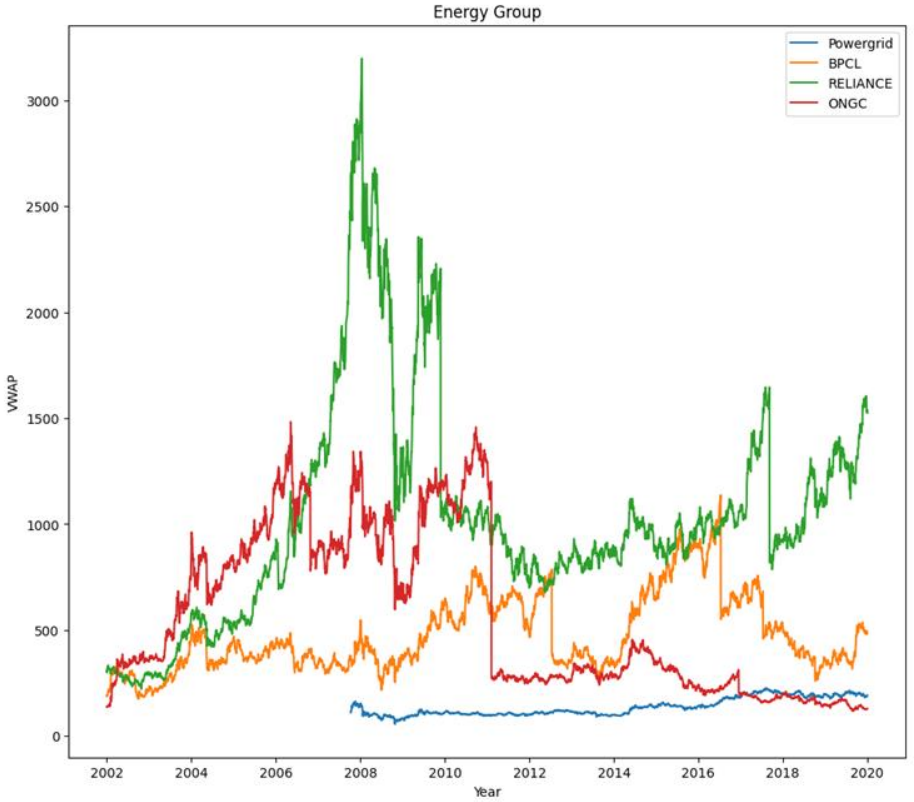
7. Variation in accuracy with Confidence(p)

Confidence (p)	***Accuracy***		
	IT vs Finance	**IT vs Energy**	**Finance vs Energy**
p>0.4	0.7698	0.4052	0.6102
p>0.5	0.7698	0.4052	0.6102
p>0.6	0.6581	0.3735	0.5526
p>0.7	0.4754	0.3670	0.5191
p>0.8	0.4665	0.3649	0.5024
p>0.9	0.4629	0.3632	0.4991

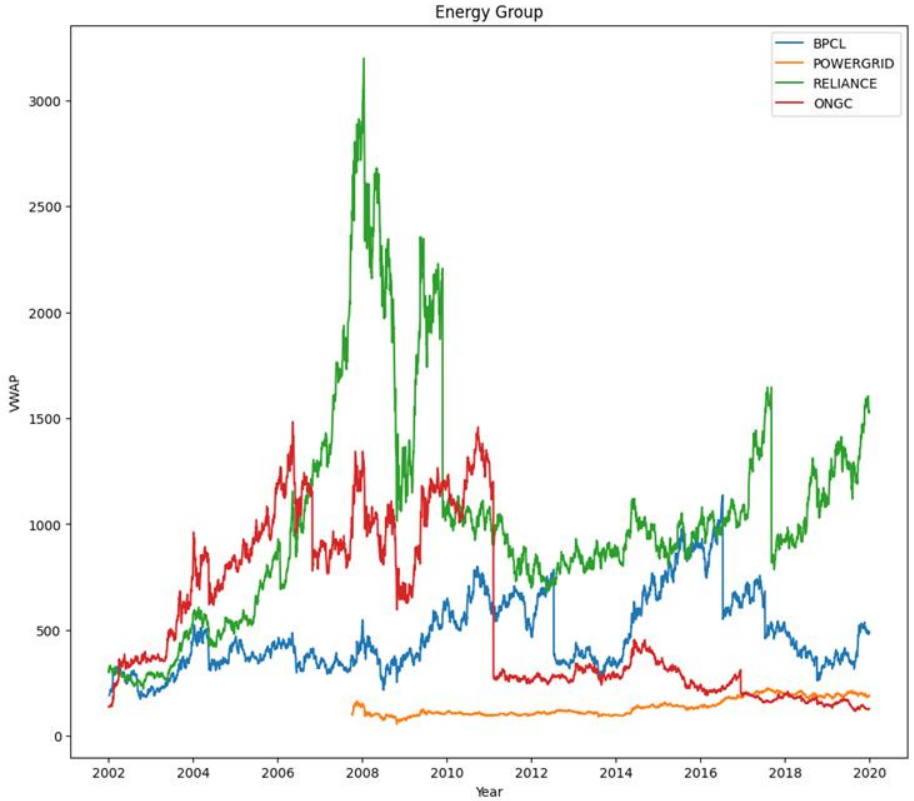


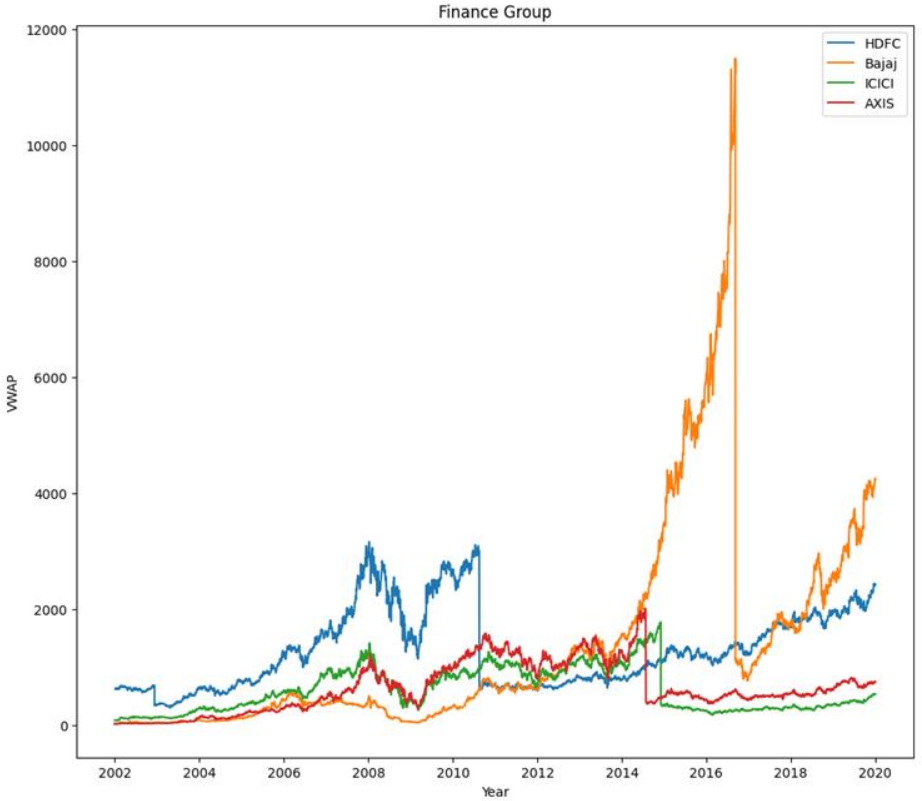
7. Reasons for Variations in Results





7. Reasons for Variations in Results





8. References

1. W. Fan and I. Davidson. On sample selection bias and its efficient correction via model averaging and unlabeled examples. In Proc. of SDM'07,2007, https://epubs.siam.org/doi/abs/10.1137/1.9781611972771.29

2. J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Sch¨olkopf. Correcting sample selection bias by unlabeled data. In Proc. of NIPS' 06, pages 601–608. 2007, https://proceedings.neurips.cc/paper/2006/hash/a2186aa7c086b46ad4e8bf81e2a3a19b-Abstract.html

3. H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference, 90(2):227–244, 2000, https://www.sciencedirect.com/science/article/pii/S0378375800001154

4. Saigal, P., & Khanna, V. Multi-category news classification using Support Vector Machine based classifiers. SN Applied Sciences, 2(3), 458., (2020), https://link.springer.com/article/10.1007/s42452-020-2266-6

5. Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2007, August). Co-clustering based classification for out-of-domain documents. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 210- 219).,2007, https://dl.acm.org/doi/abs/10.1145/1281192.1281218

6. Ieracitano, C., Adeel, A., Gogate, M., Dashtipour, K., Morabito, F. C., Larijani, H., ... & Hussain, A. (2018). Statistical analysis driven optimized deep learning system for intrusion detection. In Advances in Brain Inspired Cognitive Systems: 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings 9 (pp. 759-769). Springer International Publishing.,2018, https://researchonline.gcu.ac.uk/ws/portalfiles/portal/26504106/BICS_2018_paper_65.pdf

7. Danny Vanpoucke, https://github.com/DannyVanpoucke/LSSVMlib/blob/master/LSSVMlib/LSSVMRegression.py, Accessed on 20th March, 2024.

8. Alind Gupta, https://www.geeksforgeeks.org/ml-spectral-clustering/, accessed in March, 2024

9. Stack Overflow, https://stackoverflow.com/questions/68384427/calculate-rolling-mean-max-min-std-of-time-series-pandas-dataframe, accessed in May, 2024